Developing a Distributed Data Dictionary Service

Jim U'Ren
KM Standards Group Lead

October 9, 2000

Draft

Abstract

In an interconnected world, applications and the datasets they create and maintain need to interconnect in much the same way that computers are connected via the Internet: i.e. through the use of well established standards-based technologies. Data dictionaries are the keys to these datasets; they outline meaning and structure of the information contained within them but what is currently missing is a mechanism that allows data dictionaries to be linked to each other. Linking data dictionaries will enable the linking of datasets and applications. This paper will explore the use of the Lightweight Directory Access Protocol (LDAP) using the ISO 11179 Data Dictionary Schema as a mechanism for standardizing the structure and communication links between data dictionaries.

Background

Data Dictionaries are not a new concept to the world of computing. They have been around since the beginning of the age of computing. Typically, every time an application is developed, a data dictionary is created and used by the application. But not all data dictionaries are created equal (just as all datasets are not created equal). Some applications create a data dictionary automatically; others require a data dictionary to be created manually; others use third-party

In general, a data dictionary is a repository for information about a database, where a database consists of one or more tables. It provides a way for managing all of the data elements that make up an application. It can be thought of as a collection of meta-data about the database. e.g. filenames, field names, data types, pedigree, classifications, aliases, references, legal field lengths, legal values, language used, publisher, create date, last-modified date, version, definitions, contact information, qualifiers, units of measure, relationships, legal rights, etc.

A data dictionary is a repository for information about a database, where a database consists of one or more tables.

It provides a way for managing all of the data elements that make up an application.

It provides a complete description of the data elements in an application.

It provides the necessary information to rebuild tables and indexes, track field definitions and purposes, and identify design flaws before they become a maintenance nightmare.

[Steve Hughes - PDS Data Dictionary Case Study - automatic validation of data at time of input]

Uses of a Data Dictionary

A documentation aid for programmers and systems analysts

A security system for the database administrator

An audit trail for accountants because it shows input sources, programs that modify or use certain data items, and output reports.

A documentation tool for accountants who may need to trace data paths in the design of a new computer system

An aid in the investigating and documenting of internal control procedures

Data dictionaries are published in a number of ways:

- Off-line database (SQL, MS Access, Excel Spreadsheet, etc)
- On-line database (HTML web page, CGI access to database)
- Publish hardcopy

In the real world, datasets are often highly dynamic and their corresponding data dictionaries should be equally dynamic. Data dictionaries can be generated manually or through automated mechanisms run against existing databases.

Despite countless successful implementation of domain specific data dictionaries, there are a number of fundamental problems that need to be addressed to take data dictionaries to the next level.

The Problem

No piece of information is an island unto itself. Every piece of information has relationships with one or more pieces of information.

The problems that need to be addressed are:

- data dictionaries are not readily available

- when available, they use incompatible data structures

- the term data dictionary means different things to different classes of users

- users and applications need to be able to quickly determine the similarities and differences between data sets.

- schemas are often distributed with applications and can go obsolete as standards are revised and updated.

The fundamental problem that must be addressed is that data dictionaries are not available on-line in using a consistent structure and access mechanisms.

Another problem that must be addressed is clarifying terms used by data dictionary community. The concept of data dictionary means different things to different people. To some it a vocabulary list of terms used in an application. To others it is a collection of the descriptions of each field of a dataset. To others, it is the relationship of the field to each other and the relationships to fields in other datasets. In fact, a data dictionary is all of these things.

There is a general need for a service that will help users and applications quickly learn the similarities and differences between data sets.

A Solution

A proposed solution is to use established standards to provide common access and common data structures: the Internet standard, LDAP (Lightweight Directory Access Protocol), to provide the common access and the ISO 11179 meta-data schema standard to provide the common data structures.

LDAP (Lightweight Directory Access Protocol) is a standardized directory service developed at the University of Michigan. In this context, a directory is like a database, but tends to contain more descriptive, attribute-value-based information. The information in a directory is generally read much more often than it is written. As a result, directories don't usually implement the more complicated transaction or roll-back schemes common to SQL databases used for doing high-volume complex updates. Directory updates are typically simple all-or-nothing changes, if they are allowed at all. Directories are tuned to give quick-response to high-volume lookup or search operations. They may have the ability to replicate information widely in order to increase availability and reliability, while reducing response time. When directory information is replicated, temporary inconsistencies between the replicas may be OK, as long as they get in sync eventually.

The term service in this context follows the Internet model of a service. That is, a service is a distributed online capability based on a standard communication protocol and data structure e.g. HTTP service, FTP service, database service., Internet Service Provider, etc. Each of these examples share the following: they provide standard interfaces to on-line information, they are generally product independent in that more that one product can be use, there is a client server model used. A service is instantiated on the Net with a server software that complies with well established Internet standards. Note: the term service in this context should not to be confused with a "for pay commercial service" or service bureau.

There are many different ways to provide a directory service. Different methods allow different

kinds of information to be stored in the directory, place different requirements on how that information can be referenced, queried and updated, how it is protected from unauthorized access, etc. Some directory services are local, providing service to a restricted context (e.g., the finger service on a single machine). Other services are global, providing service to a much broader context (e.g., the entire Internet). Global services are usually distributed, meaning that the data they contain is spread across many machines, all of which cooperate to provide the directory service. Typically a global service defines a uniform namespace which gives the same view of the data no matter where you are in relation to the data itself.

LDAP records can be created using the ISO 11179 elements providing the basic minimums for describing the elements that make up a data dictionary.

LDAP records can include a variety of data types e.g. text, strings, numeric, images, hyperlinks, etc.

A data dictionary contains vocabulary, data elements and schemas (or DTDs)[1].

LDAP services can be linked hierarchically

Searches can be made on individual or groups of LDAP servers

An LDAP service can forward a request to another linked service.

LDAP databases can be contained in a variety of formats:

LDBM

SQL

Access Db

ISO/IEC 11179 Metadata Registry Implementation Coalition

Purpose: The 11179 Metadata Registry Consortium is a forum for information exchange on the implementation of metadata registries based on the ISO/IEC-11179 Specification and Standardization of Data Element standard. Consortium members are interested in addressing

[1] This is not to confuse schemas and DTDs - - they are very different animals. What they have in common is that they are structured collections of data elements with schemas generally providing more information that DTDs.

ISO/IEC-11179 reference implementations of metadata registries, influencing commercial vendors to support ISO/IEC-11179 in their tools, developing methods to support metadata exchange between metadata registries, sharing information and lessons learned on implementation approaches, being an advocate and clearinghouse for metadata registry issues, and developing partnerships to support data management across organizations.

In integral part of the proposed solution that addresses the issue of data dictionaries mean different things to different people is to break a domain data dictionary into three conceptual parts: Vocabularies, Data Elements and Schemas

The proposed solution of using LDAP as a communication protocol and ISO 11179 as a data structure standard has the following benefits:

Open – based on publicly available standard specifications that is vendor independent

Distributed – LDAP service is designed for distributed implementations; servers can be link hierarchically

Modular – domains for dictionaries can be delineated on servers; servers can range in size

Flexible – LDAP server schema can be configured as flexible as the environment requires; minimum number of attributes can be required with additional attribute value pairs added as necessary

Extensible – LDAP schemas can be extended as necessary; extensions can on local servers can be implemented without affecting the server's links to other servers

LDAP Testbed

An LDAP server has been installed and configured to test its ability to manage namespace schemas

Here is the configurations used to make additions to the OpenLDAP's SLAPD

LDAP Clients a real strength of using the LDAP protocol is there are numerous client interfaces currently available offering the flexibility and versatility for a variety of end user applications

LDAP URL

JAVA

Perl

Python

ColdFusion

LDAP can be used to wrap existing data dictionaries

Existing Data Dictionary Services can be wrapped with LDAP much the same way that x500 address services have been wrapped by LDAP. This

Commercial products exist that support wrapping existing databases with an LDAP interface.

Scenarios of the Distributed Data Dictionary Service

Terminology lookup -

(1) an end user needing to clarify use and meaning of a word used within a specific context, does a multi-domain vocabulary lookup across multiple DD services looking for the published vocabulary of the referenced domain

(2) a search engine as it is reading a document discovers a keyword list and finds a "reserved word"; the document includes a reference to a DD service where vocabulary pertaining to the document can be found; the search engine uses this vocabulary to be certain it is indexing the keywords in the right context

(3) an end user is confronted with a number of acronyns he is unfamilar with; using a local DD Service, they are able to lookup the acronyms in question and to identify alternative meanings eg. STEP stds work vs the JPL STEP project

(4) modeler references a vocabulary list associated with a data element as he is building or extending a schema

- validation of exchanged data sets - (1) a geometry model is sent from design to analysis and validation is performed using the correct schema version as refereced in the model; validation occurrs as a matter of practice before any work is done with the model

(2) a system integrator receives a AP203 model of a component to be integrated into any assembly; first step should be to run a validation routine on the model looking up using the schema specified in the Part 21 file

(3) a STEP model is checked into a PDM system which runs an automated validation routine that checks the model using the schema (located in the DD Service) that is identified in the Part 21 file.

Creating a TDP - lookup application schema (e.g. AP203, AP210, AP233, etc.) and TDP schema

(e.g. AP232, PDM schema) when generating a TDP

Data modelling (reuse)

(1) building a new schema from a published collection of data items

(2) extending a schema to solve a local problem using data elements from a published collection of data items.

Data integration -

(1) an analyst is charged to integrate data from two or more data sets; the correct schemas called out in each data set are looked up inthe DD Service to identify / map interfaces between two or more data sets

e.g. MCAD-ECAD-Cost data

- data modeling - a data modeler charged with developing a information model for a new application uses data elements published in several DD Services (much like a parts library) ensuring that the new information model will have compatible interfaces with data sets that share the same data elements.

Publishing

(1) domain owner publishes vocabulary

(2) information modeler publishes xml schema

(3) information modeler updates xml schema

(4) information modeler reuses existing xml schema to build a new xml schema

- B2B transaction - business A sends out a bid to business B, C and D; bid sent electronically as an XML object with a referenced DTD; business C reviews bid and has questions regarding the definitions of several terms; by following the link to the referenced DTD which referenced the related vocabulary, business C is able to quickly establish the underlying meaning of the bid without having to guess and respond with a proposal that is on target; businesses B and D are clueless because they don't know what a DTD is - - just kidding :-).

The Process and Challenges of Growing a Distributed a Data Dictionary Service

How a data dictionary is developed with lead to how it is operated and maintained

Implementation Considerations

LDAP DN Naming conventions

Service Linkages – hierarchies and LDAP DN considerations.

Schema checking vs flexible schemas – what attributes should be required

This service must operate within constraints of any given copyright limitations. LDAP Security mechanisms can be employed as required.

What is next?

Test LDAP DD Service connectivity

Test LDAP Client Interfaces

Test DN naming convention to enabling search routines and connectivity

Conclusions

Combining the LDAP protocol with the ISO 11179 meta-data schema leverages and combines two established standards.

The purpose of the DD Service is not to create uniformity but to enhance communication for only through communication can we determine what we have in common and where features and content are unique.

References

CMP TechEncyclopedia http://www.techweb.com/encyclopedia/

ISO/IEC 11179 – a Draft International Standard on the Specification and Standardization of Data Elements; from the Joint Technical Committee 1, ISO/IEC JTC1

ISO 10303-1

[ISO Vocabulary Standards from B. Wenzel's slides]

RFC 2251, Lightweight Directory Access Protocol (v3), M. Wahl, T. Howes, S. Kille, December 1997

RFC 2307 An Approach for Using LDAP as a Network Information Service. L. Howard. March 1998.

STEP Glossary from PDES Inc., a STEP industry consortium, http://jau.jpl.nasa.gov/step/GLOSSARY_STEP-terms-PDES-Inc.htm

Tim Berners-Lee, Glossary from Weaving the Web http://www.w3.org/People/Berners-Lee/Weaving/glossary.html

W3C DOM Glossary http://www.w3.org/TR/PR-DOM-Level-1/glossary.html

Appendix A - Glossary

attribute - A characteristic of an object or entity ISO11179

data element - A unit of data for which the definition, identification, representation, and permissiblevalues are specified by means of a set of attributes ISO11179

DN – [Distinguished Name] -

DTD – [Document Type Definition] - a DTD is a metadocument containing information about how a given set of SGML tags can be used. In the XML world this role will be taken by a schema. Sometimes, but arguably, "document type definition." Tim Berners-Lee

IEC – [International Electrotechnical Commission]

ISO – [International Standards Organization]

IETF – [Internet Engineering Task Force]

LDAP – [Lightweight Directory Access Protocol]

namespace - A name or group of names that are defined according to some naming convention. A flat namespace uses a single, unique name for every device. For example, a small Windows (NetBIOS) network requires a different, made-up name for each computer and printer. The Internet uses a hierarchical namespace that partitions the names into categories known as top level domains such as .com, .edu and .gov, etc., which are at the top of the hierarchy. See Internet domain names and XML namespace.CMP

RFC – [Request for Comment]

schema – The definition of an entire database. It defines the structure and the type of contents that each data element within the structure can contain. Schemas are often designed with visual modeling tools that automatically create the SQL code necessary to define the table structures. CMP

semantics - the branch of linguistic science which deals with the meaning of words Webster

vocabulary – a list or collection of words or of words and phrases alphabetically arranged and explained or defined

RFC

Appendix B - Configuration of the prototype Distributed Data Dictionary Service

The slapd.conf file looks like:

```
# See slapd.conf(5) for details on configuration options.
# This file should NOT be world readable.
#
include         /usr/local/etc/openldap/slapd.at.conf
include         /usr/local/etc/openldap/slapd.oc.conf
schemacheck     off
#referral       ldap://ldap.itd.umich.edu

pidfile         /usr/local/var/slapd.pid
argsfile        /usr/local/var/slapd.args

########################################################################
# ldbm database definitions
########################################################################

database        ldbm
suffix      "dc=JPL, dc=US"
directory       /usr/tmp
rootdn          "cn=root, dc=JPL, ,dc=NASA, dc=gov"
rootpw          xxxxxx
# cleartext passwords, especially for the rootdn, should
# be avoid.  See slapd.conf(5) for details.
```

The LDIF input file (input-record.ldif) looks like:

```
dn: cn=description, dc=JPL, dc=US
cn: description
o: Jet Propulsion Laboratory
objectclass: organization
objectclass: dcObject
```

Here is what the ldapadd command line looks like with the verbose (-v) feedback switched on:

```
# ldapadd -v -D "cn=root, dc=JPL, dc=US" -W < test4.ldif
Enter LDAP Password:
add cn:
        description
add o:
        Jet Propulsion Laboratory
add objectclass:
        organization
        dcObject
adding new entry cn=description, dc=JPL, dc=US
modify complete
```

#